Remarks

Reconsideration of this application is now being requested. Claims 1-20 have been canceled. Claims 21-39 have been added.

The office action alleged that a substitute specification including the claims, abstract, and title is required pursuant to 37 CFR 1.125(a) because the nature of the specification currently of record renders it difficult to consider the application. A substitute specification is being provided herein. No new matter has been added to the substitute specification.

Claims 1-20 were objected to under 37 CFR 1.75(a) for various informalities. Claims 1-20 have been canceled. Claims 21-39 have been added. Support for claims 21-39 can be found at page 13, line 1 through page 16, line 11 of the original specification and clean version of the substitute specification.

Claims 1-20 were rejected for one or more reasons. Claims 1-20 have been canceled. Claims 21-39 have been added. Support for claims 21-39 can be found at page 13, line 1 through page 16, line 11 of the original specification and clean version of the substitute specification.

By Stephen M Gurey

Stephen M. Gurey
Attorney for the Applicant
Reg. No. 27,336
(973)-386-8252

Att.
Specification – Marked Up Version
Specification – Clean Version

Date: **14 August 2006**

Docket Administrator (Room 3J-219)
Lucent Technologies Inc.
101 Crawfords Corner Road
Holmdel, NJ 07733-3030

OIPE IAP25 AUG 16 2006 PATENT & TRADEMARK OFFICE

METHOD OF REFLECTING TIME/LANGUAGE DISTORTION IN OBJECTIVE SPEECH QUALITY ASSESSMENT

Field of the Invention

The present invention relates generally to communications systems and, in particular, to speech quality assessment.

Background of the Related Art

Performance of a wireless communication system can be measured, among other things, in terms of speech quality. In the current art, there are two techniques of speech quality assessment. The first technique is a subjective technique (hereinafter referred to as "subjective speech quality assessment"). In subjective speech quality assessment, human listeners are typically used to rate the speech quality of processed speech, wherein processed speech is a transmitted speech signal which has been processed at the receiver. This technique is subjective because it is based on the perception of the individual human, and human assessment of speech quality by native listeners, i.e., people that speak the language of the speech material being presented or listened, typically takes into account language effects. Studies have shown that a listener's knowledge of language affects the scores in subjective listening tests. Scores given by native listeners were lower in subjective listening tests compared to scores given by non-native listeners when language information in speech is defect, i.e., mute. In a normal telephone conversation, the listener is often a native listener. Thus, it is preferable to use native listeners for subjective speech quality assessment in order to emulate typical conditions. Subjective speech quality assessment techniques provide a good assessment of speech quality but can be expensive and time consuming.

The second technique is an objective technique (hereinafter referred to as "objective speech quality assessment"). Objective speech quality assessment is not based on the perception of the individual human. Some objective speech quality assessment techniques are based on known source speech or reconstructed source speech estimated from processed speech. Other objective speech quality assessment techniques are not based on known source speech but on processed speech only. These latter techniques are referred to herein as "single-ended objective speech quality assessment techniques" and are often used when known source speech or reconstructed source speech are unavailable.

Current single-ended objective speech quality assessment techniques, however, do not provide as good an assessment of speech quality compared to subjective speech quality assessment techniques. One reason why current single-ended objective speech quality assessment techniques are not as good as subjective speech quality assessment techniques is because the former techniques do not account for language effects. Current single-ended objective speech quality assessment techniques have been unable to account for language effects in its speech assessment.

Accordingly, there exists a need for a single-ended objective speech quality assessment technique which accounts for language effects in assessing speech quality.

Summary of the Invention

The present invention is an objective speech quality assessment technique that reflects the impact of distortions which can dominate overall speech quality assessment by modeling the impact of such distortions on subjective speech quality assessment, thereby, accounting for language effects in objective speech quality assessment. In one embodiment, the objective speech quality assessment technique of the present invention comprises the steps of detecting distortions in an interval of speech activity using envelope information, and modifying an objective speech quality assessment value associated with the speech activity to reflect the impact of the distortions on subjective speech quality assessment. In one embodiment, the objective speech quality assessment technique also distinguish types of distortions, such as short bursts, abrupt stops and abrupt starts, and modifies the objective speech quality assessment values to reflect the different impacts of each type of distortion on subjective speech quality assessment.

Brief Description of the Drawings

The features, aspects, and advantages of the present invention will become better understood with regard to the following description, appended claims, and accompanying drawings where:

Fig. 1 depicts a flowchart illustrating an objective speech quality assessment technique accounting for language effects in accordance with one embodiment of the present invention;

Fig. 2 depicts a flowchart illustrating a voice activity detector (VAD) which detects voice activity by examining envelope information associated with the speech signal in accordance with one embodiment of the present invention;

Fig. 3 depicts an example VAD activity diagram illustrating intervals T and G of speech and non-speech activities, respectively;

Fig. 4 depicts a flowchart illustrating an embodiment for determining whether speech activity is a short burst or impulsive noise and for modifying objective speech frame quality assessment $v_s(m)$ when a short burst or impulsive noise is determined;

Fig. 5 depicts a flowchart illustrating an embodiment for determining whether speech activity has an abrupt stop or mute and for modifying objective speech frame quality assessment $v_s(m)$ when it is determined that such speech activity has an abrupt stop or mute; and

Fig. 6 depicts a flowchart illustrating an embodiment for determining whether speech activity has an abrupt start and for modifying objective speech frame quality assessment $v_s(m)$ when it is determined that such speech activity has an abrupt start.

Detailed Description

The present invention is an objective speech quality assessment technique that reflects the impact of distortions which can dominate overall speech quality assessment by modeling the impact of such distortions on subjective speech quality assessment, thereby, accounting for language effects in objective speech quality assessment.

Fig. 1 depicts a flowchart 100 illustrating an objective speech quality assessment technique accounting language effects in accordance with one embodiment of the present invention. In step 102, speech signal *s(n)* is processed to determine objective speech frame quality assessment $v_s(m)$, i.e., objective quality of speech at frame *m*. In one embodiment, each frame *m* corresponds to a 64 ms interval. The manner of processing a speech signal *s(n)* to obtain objective speech frame quality assessment $v_s(m)$ (which do not account for language effects) is well-known in the art. One example of such processing is described in co-pending application serial number 10/186,862, entitled "Compensation Of Utterance-Dependent Articulation For Speech Quality Assessment", filed on July 01, 2002 by inventor Doh-Suk Kim, ~~attached herein as Appendix A~~ which is being incorporated herein by reference.

In step 105, speech signal *s(n)* is analyzed for voice activity by, for example, a voice activity detector (VAD). VADs are well-known in the art. Fig. 2 depicts a flowchart 200 illustrating a VAD which detects voice activity by examining envelope information associated with the speech signal in accordance with one embodiment of the present invention. In step 205, envelope signals $\gamma_k(n)$ are summed up for all cochlear channels *k* to form summed envelope signal $\gamma(n)$ in accordance with equation (1):

$$\gamma(n) = \sum_{k=1}^{N_{cb}} \gamma_k(n) \qquad \text{equation (1)}$$

where $\gamma_k(n) = \sqrt{s_k^2(n) + \hat{s}_k^2(n)}$, *n* represents a time index, N_{cb} represents a total number of critical bands, $s_k(n)$ represents the output of speech signal *s(n)* through cochlear channel *k*, i.e., $s_k(n) = s(n) * h_k(n)$, and $\hat{s}_k(n)$ is the Hilbert transform of $s_k(n)$.

In step 210, a frame envelope *e(l)* is computed every 2 ms by multiplying summed envelope signal $\gamma(n)$ with a 4 ms Hamming window *w(n)* in accordance with equation (2):

$$e(l) = log\left[\sum_{n=0}^{31} \gamma^{(l)}(n) w(n) + 1\right] \qquad \text{equation (2)}$$

where $\gamma^{(l)}(n)$ is the 2 ms *l*-th frame signal of the summed envelope signal $\gamma(n)$. It should be understood that the durations of the frame envelope *e(l)* and Hamming window *w(n)* are merely illustrative and that other durations are possible. In step 215, a flooring operation is applied to frame envelope *e(l)* in accordance with equation (3).

$$e(l) = \begin{cases} e(l) & \text{if } e(l) > 5 \\ 5 & \text{otherwise} \end{cases} \qquad \text{equation (3)}$$

In step 220, time derivative $\Delta e(l)$ of floored frame envelope $e(l)$ is obtained in accordance with equation (4).

$$\Delta e(l) = \frac{\sum_{j=-3}^{3} j e(l-j)}{\sum_{j=-3}^{3} j^2} \qquad \text{equation (4)}$$

where $-3 \leq j \leq 3$.

In step 225, voice activity detection is performed in accordance with equation (5).

$$vad(l) = \begin{cases} 1 & \text{if } e(l) > 5 \\ 0 & \text{otherwise} \end{cases} \qquad \text{equation (5)}$$

In step 230, the result of equation (5), i.e., *vad(l)*, can then be refined based on the duration of 1's and 0's in the output. For example, if the duration of 0's in *vad(l)* is shorter than 8 ms, then *vad(l)* shall be changed to 1's for that duration. Similarly, if the duration of 1's in *vad(l)* is shorter than 8 ms, the *vad(l)* shall be changed to 0's for that duration. Fig. 3 depicts an example VAD activity diagram 30 illustrating intervals T and G of speech and non-speech activities, respectively. It should be understood that speech activities associated with intervals T may include, for example, actual speech, data or noise.

Returning to flowchart 100 of Fig. 1, upon analyzing speech signal *s(n)* for speech activity, interval T is examined to determine whether the associated speech activity corresponds to a short burst or impulsive noise in step 110. If the speech activity in interval T is determined to be a short burst or impulsive noise, then objective speech frame quality assessment $v_s(m)$ is modified in step 115 to obtain a modified objective speech frame quality assessment $\tilde{v}_s(m)$. The modified objective speech frame quality assessment $\tilde{v}_s(m)$ accounts for the effects of short burst or impulsive noise by modeling or simulating the impact of short bursts or impulsive noise on subjective speech quality assessment.

From step 115 of if in step 110 the speech activity in interval T is not determined to be a short burst or impulsive noise, then flowchart 100 proceeds to step 120 where the speech activity in interval T is examined to determine whether it has an abrupt stop or mute. If the speech activity in interval T is determined to have an abrupt stop or mute, then objective speech frame quality assessment $v_s(m)$ is modified in step 125 to obtain a modified objective speech frame quality assessment $\tilde{v}_s(m)$. The modified objective speech frame quality assessment $\tilde{v}_s(m)$ accounts for the effects of the abrupt stop or mute by modeling or simulating the impact of an abrupt stop or mute and subsequent release on subjective speech quality assessment.

From step 125 or if in step 120 the speech activity in interval T is not determined to have an abrupt stop or mute, then flowchart 100 proceeds to step 130 where the speech activity in interval T is examined to determine whether it has an abrupt start. If the speech activity in interval T is determined to have an abrupt start, then objective speech frame quality assessment $v_s(m)$ is modified in step 135 to obtain a modified objective speech frame quality assessment $\tilde{v}_s(m)$. The objective speech frame quality assessment $v_s(m)$ accounts for the effects of the abrupt start by modeling or simulating the impact of an abrupt start on subjective speech quality assessment. From step 135 or if in step 130 the speech activity in interval T is not determined to have an abrupt start, then flowchart 100 proceeds to step 145 where the results of modifications to objective speech frame quality assessment $v_s(m)$, if any, are integrated into the original objective speech frame quality assessment $v_s(m)$ of step 102.

Techniques for determining whether speech activity is a short burst (or impulsive noise) or has an abrupt stop (or mute) or an abrupt start, i.e., steps 110, 120 and 130, along with techniques for modifying objective speech frame quality assessment $v_s(m)$, i.e., steps 115, 125 and 135, in accordance with one embodiment of the invention will now be described. Fig. 4 depicts a flowchart 400 illustrating an embodiment for determining whether speech activity is a short burst or impulsive noise and for modifying objective speech frame quality assessment $v_s(m)$ when a short burst or impulsive noise is determined. In step 405, an impulsive noise frame l_I is determined by finding a frame l in interval T_i where frame envelope $e(l)$ is maximum in accordance, for example, with equation (6):

$$l_I = \arg \max_{u_i \le l \le d_i} e(l) \qquad \text{equation (6)}$$

where u_i and d_i represents frames l at the beginning and end of interval T_i, respectively. In step 410, frame envelope $e(l_I)$ is compared to a listener threshold value indicating whether a human listener can consider the corresponding frame l_I as annoying short burst. In one embodiment, the listener threshold value is 8 -- that is, in step 410, $e(l_I)$ is checked to determine whether it is greater than 8. If frame envelope $e(l_I)$ is not greater than the listener threshold value, then in step 415 the speech activity is determined not to be a short burst or impulsive noise.

If frame envelope $e(l_I)$ is greater than the listener threshold value, then in step 420 the duration of interval T_i is checked to determine whether it satisfies both a short burst threshold value and a perception threshold value. That is, interval T_i is being checked to determine whether interval T_i is not too short to be perceived by a human listener and not too long to be categorized as a short burst. In one embodiment, if the duration of interval T_i is greater than or equal to 28 ms and less than or equal to 60 ms, i.e., $28 \le T_i \le 60$, then both of the threshold values of step 420 are satisfied. Otherwise the threshold values of step 420 are not satisfied. If the threshold values of step 420 are not satisfied, then in step 425 the speech activity is determined not to be a short burst or impulsive noise.

If the threshold values of step 420 are satisfied, then in step 430 a maximum delta frame envelope $\Delta e(l)$ is determined from the frame envelopes $e(l)$ in the one or more frames prior to the beginning of interval T_i through the first one or more frames of interval T_i and subsequently compared to an abrupt change threshold value, such as 0.25. The abrupt change threshold value representing a criteria for identifying an abrupt change in the frame envelope. In one embodiment, a maximum delta frame envelope $\Delta e(l)$ is determined from frame envelope $e(u_i-1)$, i.e., frame envelope immediately preceding interval T_i, through the frame envelope $e(u_i+5)$, i.e., fifth frame envelope in interval T_i, and compared to a threshold value of 0.25 -- that is, in step 430, it is checked to determine whether equation (7) is satisfied:

$$\max_{u_i-1 \le l \le u_i+5} \Delta e(l) > 0.25 \qquad \text{equation (7)}$$

If the maximum delta frame envelope $\Delta e(l)$ does not exceed the threshold value, then in step 435 the speech activity is determined not to be a short burst or impulsive noise.

If the maximum delta frame envelope $\Delta e(l)$ does exceed the threshold value, then in step 440 it is determined whether frame m_I would be sufficiently annoying to a human listener, where m_I corresponds to the frame m which is impacted most by impulsive noise frame l_I. In one embodiment, step 440 is achieved by determining whether a ratio of objective speech frame quality assessment $v_s(m_I)$ to modulation noise reference unit $v_q(m_I)$ exceeds a noise threshold value. Step 440 may be expressed, for example, using a noise threshold value of 1.1 and equation (8):

$$\frac{v_s(m_I)}{v_q(m_I)} < 1.1 \qquad \text{equation (8)}$$

wherein if equation (8) is satisfied, it would be determined that frame m_I has sufficient annoyance to a human listener. If it is determined that objective speech frame quality assessment $v_s(m_I)$ would be sufficiently annoying to a human listener, then in step 445 the speech activity is determined not to be a short burst or impulsive noise.

If it is determined that objective speech frame quality assessment $v_s(m_I)$ would not be sufficiently annoying to a human listener, then in step 450 conditions related to the durations of intervals $G_{i-1,i}$, $G_{i,i+1}$, T_{i-1} and/or T_{i+1} satisfying certain minimum or maximum duration threshold values are checked to verify that it belongs to human speech. In one embodiment, the conditions of step 450 are expressed as equations (9) and (10).

$$G_{i-1,i} < 180\text{ ms and } G_{i,i+1} > 40\text{ ms and } T_{i-1} > 50\text{ ms} \qquad \text{equation (9)}$$

$$G_{i-1,i} > 40\text{ ms and } G_{i,i+1} < 100\text{ ms and } T_{i+1} > 60\text{ ms} \qquad \text{equation (10)}$$

If any of these equations or conditions are satisfied, then in step 455 the speech activity is determined not to be a short burst or impulsive noise. Rather the speech activity is determined to be natural speech. It should be understood that the minimum and maximum duration threshold values used in equations (9) and (10) are merely illustrative and may be different.

If none of the conditions in step 450 are satisfied, then in step 460 objective speech frame quality assessment $v_s(m)$ is modified in accordance with equation 11:

$$\tilde{v}_s(m) = \frac{v_s(m)}{1+\exp\left[-8.2(m-m_I)/e(l_I)-10\right]} \qquad \text{equation (11)}$$

Fig. 5 depicts a flowchart 500 illustrating an embodiment for determining whether speech activity has an abrupt stop or mute and for modifying objective speech frame quality assessment $v_s(m)$ when it is determined that such speech activity has an abrupt stop or mute. In step 505, abrupt stop frame l_M is determined. The abrupt stop frame l_M is determined by first finding negative peaks of delta frame envelope $\Delta e(l)$ in the speech activity using all frames l in interval T_i. Delta frame envelope $\Delta e(l)$ has a negative peak at l if $\Delta e(l) < \Delta e(l+j)$ for $3 \leq j \leq 3$. Upon finding the negative peaks, abrupt stop frame l_M is determined as the minimum of the negative peaks of delta frame envelopes $\Delta e(l)$. In step 510, delta frame envelope $\Delta e(l_M)$ is checked to determined whether an abrupt stop threshold value is satisfied. The abrupt stop threshold representing a criteria for determining whether there was sufficient negative change in frame envelope from one frame l to another frame $l+1$ to be considered an abrupt stop. In one embodiment, the abrupt stop threshold value is -0.56 and step 510 may be expressed as equation (12):

$$\Delta e(l_M) < -0.56 \qquad \text{equation (12)}$$

If delta frame envelope $\Delta e(l_M)$ does not satisfy the abrupt stop threshold value, then in step 515 the speech activity is determined not to have an abrupt stop or mute.

If delta frame envelope $\Delta e(l_M)$ does satisfy the abrupt stop threshold value, then in step 520 interval T_i is checked to determine if the speech activity is of sufficient duration, e.g., longer than a short burst. In one embodiment, the duration of interval T_i is checked to see if it exceeds the duration threshold value, e.g., 60 ms. That is, if $T_i < 60$ ms, then the speech activity associated with interval T_i is not of sufficient duration. If the speech activity is considered not of sufficient duration, then in step 525 the speech activity is determined not to have an abrupt stop or mute.

If the speech activity is considered of sufficient duration, then in step 530 a maximum frame envelope $e(l)$ is determined for one or more frames prior to frame l_M through frame l_M or beyond and subsequently compared against a stop-energy threshold value. The stop-energy threshold value representing a criteria for determining whether a frame envelope has sufficient energy prior to muting. In one embodiment, maximum frame envelope *e(l)* is determined for frames l_M-7 through l_M and compared to a stop-energy threshold value of 9.5, i.e., $\max_{l_m-7\le l\le l_m} e(l) > 9.5$. If the maximum frame envelope *e(l)* does not satisfy the stop-energy threshold value, then in step 535 the speech activity is determined not to have an abrupt stop or mute.

If the maximum frame envelope *e(l)* does satisfy the stop-energy threshold value, then objective speech frame quality assessment $v_s(m)$ is modified in accordance with equation 13 for several frames *m*, such as $m_M, \ldots, m_M+6$:

$$\tilde{v}_s(m) = \left|\Delta e(l_M)\right|\left[\frac{6}{1+\exp\left[-2(m-m_M-3\right]}-6\right] \qquad \text{equation (13)}$$

where m_M corresponds to the frame *m* which is impacted most by abrupt stop frame l_M.

Fig. 6 depicts a flowchart 600 illustrating an embodiment for determining whether speech activity has an abrupt start and for modifying objective speech frame quality assessment $v_s(m)$ when it is determined that such speech activity has an abrupt start. In step 605, abrupt start frame l_S is determined. The abrupt start frame l_S is determined by first finding positive peaks of delta frame envelope $\Delta e(l)$ in the speech activity using all frames *l* in interval T_i. Delta frame envelope $\Delta e(l)$ has a positive peak at *l* if $\Delta e(l) > \Delta e(l+j)$ for $3 \le j \le 3$. Upon finding the positive peaks, abrupt start frame l_S is determined as the maximum of the positive peaks of delta frame envelopes $\Delta e(l)$. In step 610, delta frame envelope $\Delta e(l_S)$ is checked to determined whether an abrupt start threshold value is satisfied. The abrupt start threshold representing a criteria for determining whether there was sufficient positive change in frame envelope from one frame *l* to another frame *l*+1 to be considered an abrupt start. In one embodiment, the abrupt stop threshold value is 0.9 and step 610 may be expressed as equation (14):

$$\Delta e(l_S) > 0.9 \qquad \text{equation (14)}$$

If delta frame envelope $\Delta e(l_S)$ does not satisfy the abrupt start threshold value, then in step 615 the speech activity is determined not to have an abrupt start.

If delta frame envelope $\Delta e(l_S)$ does satisfy the abrupt start threshold value, then in step 620 interval T_i is checked to determined if the speech activity is of sufficient duration, e.g., longer than a short burst. In one embodiment, the duration of interval T_i is checked to see if it exceeds the short burst threshold value, e.g., 60 ms. That is, if $T_i < 60$ ms, then the speech activity associated with interval T_i is not of sufficient duration. If the speech activity is not of sufficient duration, then in step 625 the speech activity is determined not to have an abrupt start.

If the speech activity is of sufficient duration, then in step 630 a maximum frame envelope $e(l)$ is determined for frame l_S or prior through one or more frames after frame l_S and subsequently compared against a start-energy threshold value. The start-energy threshold value representing a criteria for determining whether a frame envelope has sufficient energy. In one embodiment, maximum frame envelope $e(l)$ is determined for frames l_S through l_S+7 and compared to a start-energy threshold value of 12, i.e., $\max_{l_S \le l \le l_S+7} e(l) < 12$. If the maximum frame envelope $e(l)$ does not satisfy the start-energy threshold value, then in step 635 the speech activity is determined not to have an abrupt start.

If the maximum frame envelope $e(l)$ does satisfy the start-energy threshold value, then objective speech frame quality assessment $v_s(m)$ is modified in accordance with equation 16 for several frames m, such as $m_M, \ldots, m_M+6$:

$$\tilde{v}_s(m) = \frac{v_s(m)}{1+\exp\left[-0.4(m-m_S)/\Delta e(l_S)-10\right]} \qquad \text{equation (16)}$$

where m_S corresponds to the frame m which is impacted most by abrupt start frame l_S. It should be understood that the values used in equations (11), (13) and (16) were derived empirically. Other values are possible. Thus, the present invention should not be limited to those specific values.

Note that upon determining modified objective speech frame quality assessment $\tilde{v}_s(m)$, the integration performed in step 145 may be achieved using equation (17):

$$v_s(m) = \min(v_{s,I}(m), v_{s,M}(m), v_{s,S}(m)) \qquad \text{equation (17)}$$

where $v_{s,I}(m)$, $v_{s,M}(m)$ and $v_{s,S}(m)$ correspond to the modified objective speech frame quality assessment $\tilde{v}_s(m)$ of equations 11, 13 and 16, respectively.

Although the present invention has been described in considerable detail with reference to certain embodiments, other versions are possible. For example, the orders of the steps in the flowcharts may be re-arranged, or some steps (or criteria) may be deleted from or added to the flowcharts. Therefore, the spirit and scope of the present invention should not be limited to the description of the embodiments contained herein. It should also be understood to those skilled in the art that the present invention may be implemented either as hardware or software incorporated into some type of processor.

Claims

I claim:

1. A method for objectively assessing speech quality comprising the steps of:

    detecting distortions in an interval of speech activity using envelope information; and

    modifying an objective speech quality assessment value associated with the speech activity to reflect the impact of the distortions on subjective speech quality assessment.

2. The method of claim 1, wherein the step of modifying includes the step of determining the objective speech quality assessment values for the speech activity.

3. The method of claim 1, wherein the distortions being detected are impulsive noise, abrupt stop or abrupt start.

4. The method of claim 1, wherein the step of detecting includes the step of determining a distortion type.

5. The method of claim 4, wherein the distortion type is determined to be impulsive noise if the envelope information indicates that the speech activity can be perceived by a human listener to be noise and if the interval is of a duration long enough to be perceived by a human listener but not too long for a short burst.

6. The method of claim 4, wherein the distortion type is determined to be impulsive noise if the envelope information indicates that the speech activity can be perceived by a human listener to be noise, if a ratio of the objective speech quality assessment value to a modulation noise reference unit indicates a human listener would perceive annoying noise, and if the interval is of a duration long enough to be perceived by a human listener but not too long for a short burst.

7. The method of claim 4, wherein the objective quality assessment value associated with the speech activity is modified in accordance with the following equation to obtain a modified objective quality assessment value if the distortion type is impulsive noise:

$$\tilde{v}_s(m) = \frac{v_s(m)}{1+\exp\left[-8.2(m-m_I)/e(l_I)-10\right]}$$

where $v_s(m)$ is the objective quality assessment value and $\tilde{v}_s(m)$ is the modified objective quality assessment value.

8. The method of claim 4, wherein the distortion type is determined to be abrupt stop if the envelope information indicates that there was an sufficient negative change in frame energy from one frame to another to be considered an abrupt stop and if the interval is of a duration longer than a short burst.

9. The method of claim 4, wherein the distortion type is determined to be abrupt stop if the envelope information indicates that a maximum frame envelope had sufficient energy prior to ending the interval, and if the interval is of a duration longer than a short burst.

10. The method of claim 4, wherein the objective quality assessment value associated with the speech activity is modified in accordance with the following equation to obtain a modified objective quality assessment value if the distortion type is impulsive noise:

$$\tilde{v}_s(m) = \left|\Delta e(l_M)\right|\left[\frac{6}{1+\exp\left[-2(m-m_M-3\right]}-6\right]$$

where $v_s(m)$ is the objective quality assessment value and $\tilde{v}_s(m)$ is the modified objective quality assessment value.

11. The method of claim 4, wherein the distortion type is determined to be abrupt start if the envelope information indicates that there was an sufficient positive change

in frame energy from one frame to another to be considered an abrupt start and if the interval is of a duration longer than a short burst.

12. The method of claim 4, wherein the distortion type is determined to be abrupt stop if the envelope information indicates that a maximum frame envelope had sufficient energy towards a beginning of the interval, and if the interval is of a duration longer than a short burst.

13. The method of claim 4, wherein the objective quality assessment value associated with the speech activity is modified in accordance with the following equation to obtain a modified objective quality assessment value if the distortion type is impulsive noise:

$$\tilde{v}_s(m) = \frac{v_s(m)}{1 + \exp\left[-0.4(m - m_S)/\Delta e(l_S) - 10\right]}$$

where $v_s(m)$ is the objective quality assessment value and $\tilde{v}_s(m)$ is the modified objective quality assessment value.

14. The method of claim 1 comprising the additional step of:

prior to the step of detecting, determining the interval of speech activity using the envelope information.

15. An objective speech quality assessment system comprising:

means for detecting distortions in an interval of speech activity using envelope information; and

means for modifying an objective speech quality assessment value associated with the speech activity to reflect the impact of the distortions on subjective speech quality assessment.

16. The objective speech quality assessment system of claim 15, wherein the means for modifying includes a means for determining the objective speech quality assessment values without accounting for distortions for the speech activity.

17. The objective speech quality assessment system of claim 15, wherein the distortions being detected are impulsive noise, abrupt stop or abrupt start.

18. The objective speech quality assessment system of claim 15, wherein the means for detecting includes a means for determining a distortion type.

19. The objective speech quality assessment system of claim 18, wherein the means for detecting includes a voice activity detector for detecting intervals of speech activity, wherein the means for determining a distortion type examines intervals of speech activities detected by the voice activity detector.

Abstract of the Disclosure

Disclosed is an objective speech quality assessment technique that reflects the impact of distortions which can dominate overall speech quality assessment by modeling the impact of such distortions on subjective speech quality assessment, thereby, accounting for language effects in objective speech quality assessment.

~~COMPENSATION FOR UTTERANCE DEPENDENT ARTICULATION FOR SPEECH QUALITY ASSESSMENT~~

~~Field of the Invention~~

~~The present invention relates generally to communications systems and, in particular, to speech quality assessment.~~

~~Background of the Related Art~~

~~Performance of a wireless communication system can be measured, among other things, in terms of speech quality. In the current art, there are two techniques of speech quality assessment. The first technique is a subjective technique (hereinafter referred to as "subjective speech quality assessment"). In subjective speech quality assessment, human listeners are used to rate the speech quality of processed speech, wherein processed speech is a transmitted speech signal which has been processed at the receiver. This technique is subjective because it is based on the perception of the individual human, and human assessment of speech quality typically takes into account phonetic contents, speaking styles or individual speaker differences. Subjective speech quality assessment can be expensive and time consuming.~~

~~The second technique is an objective technique (hereinafter referred to as "objective speech quality assessment"). Objective speech quality assessment is not based on the perception of the individual human. Most objective speech quality assessment techniques are based on known source speech or reconstructed source speech estimated from processed speech. However, these objective techniques do not account for phonetic contents, speaking styles or individual speaker differences.~~

~~Accordingly, there exists a need for assessing speech quality objectively which takes into account phonetic contents, speaking styles or individual speaker differences.~~

~~Summary of the Invention~~

~~The present invention is a method for objective speech quality assessment that accounts for phonetic contents, speaking styles or individual speaker differences by~~

~~distorting speech signals under speech quality assessment. By using a distorted version of a speech signal, it is possible to compensate for different phonetic contents, different individual speakers and different speaking styles when assessing speech quality. The amount of degradation in the objective speech quality assessment by distorting the speech signal is maintained similarly for different speech signals, especially when the amount of distortion of the distorted version of speech signal is severe. Objective speech quality assessment for the distorted speech signal and the original undistorted speech signal are compared to obtain a speech quality assessment compensated for utterance dependent articulation. In one embodiment, the comparison corresponds to a difference between the objective speech quality assessments for the distorted and undistorted speech signals.~~

~~Brief Description of the Drawings~~

~~The features, aspects, and advantages of the present invention will become better understood with regard to the following description, appended claims, and accompanying drawings where:~~

~~Fig. 1 depicts an objective speech quality assessment arrangement which compensates for utterance dependent articulation in accordance with the present invention;~~

~~Fig. 2 depicts an embodiment of an objective speech quality assessment module employing an auditory-articulatory analysis module in accordance with the present invention.;~~

~~Fig. 3 depicts a flowchart for processing, in an articulatory analysis module, the plurality of envelopes $a_i(t)$ in accordance with one embodiment of the invention; and~~

~~Fig. 4 depicts an example illustrating a modulation spectrum $A_i(m,f)$ in terms of power versus frequency.~~

~~Detailed Description~~

~~The present invention is a method for objective speech quality assessment that accounts for phonetic contents, speaking styles or individual speaker differences by distorting processed speech. Objective speech quality assessment tend to yield different~~

~~values for different speech signals which have same subjective speech quality scores. The reason these values differ is because of different distributions of spectral contents in the modulation spectral domain. By using a distorted version of a processed speech signal, it is possible to compensate for different phonetic contents, different individual speakers and different speaking styles. The amount of degradation in the objective speech quality assessment by distorting the speech signal is maintained similarly for different speech signals, especially when the distortion is severe. Objective speech quality assessment for the distorted speech signal and the original undistorted speech signal are compared to obtain a speech quality assessment compensated for utterance dependent articulation.~~

~~Fig. 1 depicts an objective speech quality assessment arrangement 10 which compensates for utterance dependent articulation in accordance with the present invention. Objective speech quality assessment arrangement 10 comprises a plurality of objective speech quality assessment modules 12, 14, a distortion module 16 and a compensation utterance-specific bias module 18. Speech signal s(t) is provided as inputs to distortion module 16 and objective speech quality assessment module 12. In distortion module 16, speech signal s(t) is distorted to produce a modulated noise reference unit (MNRU) speech signal s'(t). In other words, distortion module 16 produces a noisy version of input signal s(t). MNRU speech signal s'(t) is then provided as input to objective speech quality assessment module 14.~~

~~In objective speech quality assessment modules 12, 14, speech signal s(t) and MNRU speech signal s'(t) are processed to obtain objective speech quality assessments SQ(s(t) and SQ(s'(t)). Objective speech quality assessment modules 12, 14 are essentially identical in terms of the type of processing performed to any input speech signals. That is, if both objective speech quality assessment modules 12, 14 receive the same input speech signal, the output signals of both modules 12, 14 would be approximately identical. Note that, in other embodiments, objective speech quality assessment modules 12, 14 may process speech signals s(t) and s'(t) in a manner different from each other. Objective speech quality assessment modules are well known in the art. An example of such a module will be described later herein.~~

~~Objective speech quality assessments SQ(s(t) and SQ(s'(t)) are then compared to obtain speech quality assessment~~ ~~$SQ_{compensated}$, which compensates for~~

~~utterance dependent articulation. In one embodiment, speech quality assessment $SQ_{compensated}$ is determined using the difference between objective speech quality assessments SQ(s(t) and SQ(s'(t)). For example, $SQ_{compensated}$ is equal to SQ(s(t) minus SQ(s'(t)), or vice versa. In another embodiment, speech quality assessment $SQ_{compensated}$ is determined based on a ratio between objective speech quality assessments SQ(s(t) and SQ(s'(t)). For example,~~

~~$$SQ_{compensated} = \frac{SQ(s(t))+\mu}{SQ(s'(t))+\mu} \text{ or } SQ_{compensated} = \frac{SQ(s'(t))+\mu}{SQ(s(t))+\mu}$$~~

~~where μ is a small constant value.~~

~~As mentioned earlier, objective speech quality assessment modules 12, 14 are well known in the art. Fig. 2 depicts an embodiment 20 of an objective speech quality assessment module 12, 14 employing an auditory-articulatory analysis module in accordance with the present invention. As shown in Fig. 2, objective quality assessment module 20 comprises of cochlear filterbank 22, envelope analysis module 24 and articulatory analysis module 26. In objective quality assessment module 20, speech signal s(t) is provided as input to cochlear filterbank 22. Cochlear filterbank 22 comprises a plurality of cochlear filters $h_i(t)$ for processing speech signal s(t) in accordance with a first stage of a peripheral auditory system, where $i=1,2,\ldots,N_c$ represents a particular cochlear filter channel and N_c denotes the total number of cochlear filter channels. Specifically, cochlear filterbank 22 filters speech signal s(t) to produce a plurality of critical band signals $s_i(t)$, wherein critical band signal $s_i(t)$ is equal to $s(t)*h_i(t)$.~~

~~The plurality of critical band signals $s_i(t)$ is provided as input to envelope analysis module 24. In envelope analysis module 24, the plurality of critical band signals $s_i(t)$ is processed to obtain a plurality of envelopes $a_i(t)$, wherein $a_i(t)=\sqrt{s_i^2(t)+\hat{s}_i^2(t)}$ and $\hat{s}_i(t)$ is the Hilbert transform of $s_i(t)$.~~

~~The plurality of envelopes $a_i(t)$ is then provided as input to articulatory analysis module 26. In articulatory analysis module 26, the plurality of envelopes $a_i(t)$ is processed to obtain a speech quality assessment for speech signal s(t). Specifically, articulatory analysis module 26 does a comparison of the power associated with signals generated from the human articulatory system (hereinafter referred to as "articulation~~

~~power $P_A(m,i)$”) with the power associated with signals not generated from the human articulatory system (hereinafter referred to as “non-articulation power $P_{NA}(m,i)$”). Such comparison is then used to make a speech quality assessment.~~

~~Fig. 3 depicts a flowchart 300 for processing, in articulatory analysis module 26, the plurality of envelopes $a_i(t)$ in accordance with one embodiment of the invention. In step 310, Fourier transform is performed on frame m of each of the plurality of envelopes $a_i(t)$ to produce modulation spectrums $A_i(m,f)$, where f is frequency.~~

~~Fig. 4 depicts an example 40 illustrating modulation spectrum $A_i(m,f)$ in terms of power versus frequency. In example 40, articulation power $P_A(m,i)$ is the power associated with frequencies 2~12.5 Hz, and non-articulation power $P_{NA}(m,i)$ is the power associated with frequencies greater than 12.5 Hz. Power $P_{No}(m,i)$ associated with frequencies less than 2 Hz is the DC-component of frame m of critical band signal $a_i(t)$. In this example, articulation power $P_A(m,i)$ is chosen as the power associated with frequencies 2~12.5 Hz based on the fact that the speed of human articulation is 2~12.5 Hz, and the frequency ranges associated with articulation power $P_A(m,i)$ and non-articulation power $P_{NA}(m,i)$ (hereinafter referred to respectively as “articulation frequency range” and “non-articulation frequency range”) are adjacent, non-overlapping frequency ranges. It should be understood that, for purposes of this application, the term “articulation power $P_A(m,i)$” should not be limited to the frequency range of human articulation or the aforementioned frequency range 2~12.5 Hz. Likewise, the term “non-articulation power $P_{NA}(m,i)$” should not be limited to frequency ranges greater than the frequency range associated with articulation power $P_A(m,i)$. The non-articulation frequency range may or may not overlap with or be adjacent to the articulation frequency range. The non-articulation frequency range may also include frequencies less than the lowest frequency in the articulation frequency range, such as those associated with the DC-component of frame m of critical band signal $a_i(t)$.~~

~~In step 320, for each modulation spectrum $A_i(m,f)$, articulatory analysis module 26 performs a comparison between articulation power $P_A(m,i)$ and non-articulation power $P_{NA}(m,i)$. In this embodiment of articulatory analysis module 26, the comparison between articulation power $P_A(m,i)$ and non-articulation power $P_{NA}(m,i)$ is an~~

~~articulation-to-non-articulation ratio ANR(m,i). The ANR is defined by the following equation~~

$$ANR(m,i)=\frac{P_A(m,i)+\epsilon}{P_{NA}(m,i)+\epsilon} \qquad \text{equation (1)}$$

~~where ϵ is some small constant value. Other comparisons between articulation power $P_A(m,i)$ and non-articulation power $P_{NA}(m,i)$ are possible. For example, the comparison may be the reciprocal of equation (1), or the comparison may be a difference between articulation power $P_A(m,i)$ and non-articulation power $P_{NA}(m,i)$. For ease of discussion, the embodiment of articulatory analysis module 26 depicted by flowchart 300 will be discussed with respect to the comparison using ANR(m,i) of equation (1). This should not, however, be construed to limit the present invention in any manner.~~

~~In step 330, ANR(m,i) is used to determine local speech quality LSQ(m) for frame m. Local speech quality LSQ(m) is determined using an aggregate of the articulation-to-non-articulation ratio ANR(m,i) across all channels i and a weighing factor R(m,i) based on the DC-component power $P_{No}(m,i)$. Specifically, local speech quality LSQ(m) is determined using the following equation~~

$$LSQ(m)=\log\left[\sum_{i=1}^{N_c} ANR(m,i)R(m,i)\right] \qquad \text{equation (2)}$$

~~where~~

$$R(m,i)=\frac{\log(1+P_{No}(m,i)}{\sum_{k=1}^{Nc}\log(1+P_{No}(m,k)} \qquad \text{equation (3)}$$

~~and k is a frequency index.~~

~~In step 340, overall speech quality SQ for speech signal s(t) is determined using local speech quality LSQ(m) and a log power $P_s(m)$ for frame m. Specifically, speech quality SQ is determined using the following equation~~

$$SQ=L\{P_s(m)LSQ(m)\}_{m=1}^{T}=\left[\sum_{\substack{m=1\\ P_s>P_{th}}}^{T} P_s^{\lambda}(m)LSQ^{\lambda}(m)\right]^{1/\lambda} \qquad \text{equation (4)}$$

~~where $P_s(m) = \log\left[\sum_{t \in m} s^2(t)\right]$, L is L_p norm, T is the total number of frames in speech signal s(t), λ is any value, and P_{th} is a threshold for distinguishing between audible signals and silence. In one embodiment, λ is preferably an odd integer value.~~

~~The output of articulatory analysis module 26 is an assessment of speech quality SQ over all frames m. That is, speech quality SQ is a speech quality assessment for speech signal s(t).~~

~~Although the present invention has been described in considerable detail with reference to certain embodiments, other versions are possible. Therefore, the spirit and scope of the present invention should not be limited to the description of the embodiments contained herein.~~

~~Claims~~

~~I claim:~~

~~1. A method of assessing speech quality comprising the steps of:~~

~~determining a first and second speech quality assessment for a first and second speech signal, the first speech signal being a distorted version of the second speech signal; and~~

~~comparing the first and second speech qualities to obtain a compensated speech quality assessment.~~

~~2. The method of claim 1 comprising the additional steps of~~

~~prior to determining the first and second speech quality assessments, distorting the second speech signal to produce the first speech signal.~~

~~3. The method of claim 1, wherein the first and second speech qualities are assessed using an identical technique for objective speech quality assessment.~~

~~4. The method of claim 1, wherein the compensated speech quality assessment corresponds to a difference between the first and second speech qualities.~~

~~5. The method of claim 1, wherein the compensated speech quality assessment corresponds to a ratio between the first and second speech qualities.~~

~~6. The method of claim 1, wherein the first and second speech qualities are assessed using auditory-articulatory analysis.~~

~~7. The method of claim 1, wherein the step assessing the second or first speech quality comprises the steps of;~~

~~comparing articulation power and non-articulation power for the speech signal or distorted speech signal, wherein articulation and non-articulation powers~~

~~are powers associated with articulation and non-articulation frequencies of the speech signal or distorted speech signal; and~~

~~and assessing the second or first speech quality based on the comparison.~~

~~8. The method of claim 7, wherein the articulation frequencies are approximately 2~12.5 Hz.~~

~~9. The method of claim 7, wherein the articulation frequencies correspond approximately to a speed of human articulation.~~

~~10. The method of claim 7, wherein the non-articulation frequencies are approximately greater than the articulation frequencies.~~

~~11. The method of claim 7, wherein the comparison between the articulation power and non-articulation power is a ratio between the articulation power and non-articulation power.~~

~~12. The method of claim 10, wherein the ratio includes a denominator and numerator, the numerator including the articulation power and a small constant, the denominator including the non-articulation power plus the small constant.~~

~~13. The method of claim 7, wherein the comparison between the articulation power and non-articulation power is a difference between the articulation power and non-articulation power.~~

~~14. The method of claim 7, wherein the step of assessing the first or second speech quality includes the step of:~~

~~determining a local speech quality using the comparison.~~

~~15. The method of claim 7, wherein the local speech quality is further determined using a weighing factor based on a DC-component power.~~

~~16. The method of claim 9, wherein the first or second speech quality is determined using the local speech quality.~~

~~17. The method of claim 7, wherein the step of comparing articulation power and non-articulation power includes the step of:~~

~~performing a Fourier transform on each of a plurality of envelopes obtained from a plurality of critical band signals.~~

~~18. The method of claim 7, wherein the step of comparing articulation power and non-articulation power includes the step of:~~

~~filtering the speech signal to obtain a plurality of critical band signals.~~

~~19. The method of claim 18, wherein the step of comparing articulation power and non-articulation power includes the step of:~~

~~performing an envelope analysis on the plurality of critical band signals to obtain a plurality of modulation spectrums.~~

~~20. The method of claim 18, wherein the step of comparing articulation power and non-articulation power includes the step of:~~

~~performing a Fourier transform on each of the plurality of modulation spectrums.~~

~~Abstract of the Disclosure~~

~~A method for objective speech quality assessment that accounts for phonetic contents, speaking styles or individual speaker differences by distorting speech signals under speech quality assessment. By using a distorted version of a speech signal, it is possible to compensate for different phonetic contents, different individual speakers and different speaking styles when assessing speech quality. The amount of degradation in the objective speech quality assessment by distorting the speech signal is maintained similarly for different speech signals, especially when the amount of distortion of the distorted version of speech signal is severe. Objective speech quality assessment for the distorted speech signal and the original undistorted speech signal are compared to obtain a speech quality assessment compensated for utterance dependent articulation.~~